CHAPTER A-1 PROBABILITY AND STATISTICS

A-1.1 Key Concepts

During a risk analysis, various numbers called probabilities are estimated and used to describe our degree of belief in the likelihood of events in order to characterize the risks associated with those events. Probability is a measure of the likelihood that an event will occur. The mathematical theory of probability tells us how to apply the numbers in a logical and consistent manner. Facilitators and analysts are responsible for defining what is being estimated and ensuring that the probability estimates represent what the events they are intended to describe.

Do probabilities model a process that is random or do they describe the state of knowledge? In the case of a flood frequency curve, the answer is both. The chance of exceeding the 100-year flood elevation this year is typically modeled as a random process described by the frequency curve. The likelihood that the 100-year flood elevation will be within a particular range is typically modeled as a degree of belief described by the uncertainty about the frequency curve. This separation of the probabilities based on the source and nature of the uncertainty exists only in the risk model and not in the real world. The occurrence of floods is not necessarily random. We choose to model them as a random process because we don't have (and perhaps can't obtain) sufficient knowledge to predict long-term weather patterns at specific locations. Modeling flood occurrences as a random process might provide a more convenient or improved understanding of risk compared with attempting to model the knowledge uncertainty in weather patterns. Separation of uncertainty in a risk analysis is an important modeling decision. Probabilities associated with randomness are statements about frequency of occurrence in time or space. Probabilities associated with knowledge uncertainty are statements about our degrees of belief regarding a particular claim.

Probabilities can be estimated using a variety of techniques. Statistical estimates can be made based on past observations using empirical data. Analytical models based on physical processes and reasoning from first principles can be applied. Expert opinions can be elicited to obtain probabilities in cases where data or models are incomplete. In practice, the risk analyst should combine all of these methods when feasible to support robust probability estimates. Judgment should always be applied as an overlay to these methods to express our degree of belief in the adequacy of the data, methods, parameters, and models.

A-1.2 Deductive and Inductive Logic

Critical thinking is the application of reason to evaluate the extent to which a claim is believed to be true. It requires a "disciplined process of actively and skillfully conceptualizing, applying, analyzing, synthesizing, and/or evaluating information gathered from, or generated by, observation, experience, reflection, reasoning, or communication, as a guide to belief and action." (Scriven and Paul, 1987) All levels of decision making are impacted by critical thinking. In order to make better decisions, we must assess our beliefs and the beliefs of others. Questioning the beliefs and rationale of others is an essential element of the critical thinking process and should not be viewed as a personal criticism. The wise analyst and decision maker is one who knows how to identify and minimize errors and biases in critical thinking to support more credible decisions.

Logical arguments derive from the process of evaluating whether or not a claim is believed to be true. These arguments can be characterized as either deductive or inductive. Deductive arguments arrive at conclusions that are guaranteed to be true given certain premises. Inductive arguments support conclusions that are likely or probable based on the supporting evidence. In practice, actual truth is a challenging matter to assess. How do you know for certain that other countries in the world exist? Have you visited them? Is the map correct? Is the geography teacher correct?

Deductive arguments are valid when the conclusions necessarily follow, if the premises are assumed to be true. A valid deduction does not require the premises to actually be true. This is a potential source of disagreement among rational people because the truth of a deduction is independent of its validity. Given a premise that *all dams reduce flood risk* and that *John Rapids is a dam,* it necessarily follows that John Rapids Dam must reduce flood risk. This is a valid deduction even though we know the premise that *all dams reduce flood risk* is not true. Many navigation and hydropower dams are not designed to reduce flood risk. A deduction is sound if and only if the deduction is valid and all of its premises are actually true. In the previous example, the argument that John Rapids Dam reduces flood risk is not sound because the premise is false. However, both invalid deductions and valid deductions with unsound arguments can still have true conclusions. John Rapids Dam might reduce flood risk even though the argument is unsound. The conclusion of a deductive argument should not be automatically assumed to be untrue because of flaws in its validity or soundness.

Inductive arguments provide a framework to address issues of truth by supporting certain conclusions that are more reasonable to believe than others, but are not certain to be correct. If sand boils were observed near the levee toe during the last flood, then it is likely that sand boils will be observed during the next flood. We can't know this for certain, but our conclusion is rationally supported by the available evidence (and first principles of soil mechanics). The conclusion of an inductive argument becomes more likely to be true as more supporting evidence is obtained. As a simple example, consider a box containing 100 piezometers. Without looking inside the box, we begin to pull piezometers out. The first few piezometers we pull out are broken. At what point would you conclude that most of the remaining piezometers are also broken? Our confidence that this conclusion is true will increase as we remove and test more piezometers. We might consider other evidence to bolster our conclusion, such as observing damage to the box or witnessing the delivery person dropping the box. The more evidence we obtain, the stronger we believe in our conclusion. In practice, the amount of evidence will be limited and we must rely on the judgment and experience of experts.

Because induction is not an exact science and evidence is often limited, errors in reasoning can occur and it is possible to reach wrong conclusions. Recognizing and mitigating issues that introduce reasoning errors (e.g. overestimating the strength of evidence, overconfidence in expert judgment, group think, misplacing burden of proof, and many others) can strengthen inductive arguments. The validity of inductive conclusions must be evaluated against alternative conclusions to determine how strong they are. Decision makers often use objective standards to compare alternatives and assess whether a particular course of action is preferred over another.

In practice, both deductive and inductive arguments are necessary for a credible systematic approach to risk analysis. Deduction can provide absolute proof for a conclusion, but the premises can rarely be tested and verified to be actually true. Induction is driven by the available evidence, but proof of a theory cannot be obtained. Risk analysis requires a careful synthesis of these two logical approaches.

A-1.3 Set Theory

Many of the characteristics of a risk analysis problem can be described and modeled using sets. Set theory is a branch of mathematics that deals with the properties and relationships of collections of elements or events. Risk analysis relies on set theory to provide a logical framework for the analysis of events and the relationships between or among a collection of events.

A set is a well-defined collection of unique elements or events. The sample space for a set includes all possible outcomes of a random trial or experiment. For example, a random trial or experiment might be a levee exposed to a flood loading. The sample space for this trial might be represented by a set containing two possible events {levee breaches, levee does not breach}. The complement of an event A, includes all of the events that are not A. The complement of A can be denoted as A', A^C, or $\bar{A}$. For the levee example, the complement of {levee breaches} would be {levee does not breach}. Events are mutually exclusive when they cannot occur during the same random trial or experiment. The events {levee breaches} and {levee does not breach} are mutually exclusive because both cannot occur. Events are collectively exhaustive when at least one of the events must occur during a random trial or experiment. The collectively exhaustive events for levee performance would be {levee breaches, levee does not breach}. The union of two (or more) events, denoted by $A \cup B$, is the set that includes all outcomes that are either A or B or both. Given two potential failure modes, the union, expressed as PFM1 $\cup$ PFM2, would be the set {PFM1, PFM2, PFM1 and PFM2}. The intersection of two (or more) events, denoted by $A \cap B$ or AB, is the set of all outcomes that include both A and B. For the two potential failure mode example, the intersection, expressed as PFM1PFM2, would be the set {PFM1 and PFM2}. When the events are mutually exclusive, the intersection is an empty or null set containing no elements, {}.

Consider the following example: A flood-overtopping potential failure mode developed by a risk analysis team for an embankment dam forming a pumped storage reservoir consists of three events: A, a flood occurs; B, the reservoir elevation exceeds the available freeboard; and C, a breach occurs. In order for failure to occur by this potential failure mode, all three of these events must occur. Failure of the dam for this potential failure mode can be described as the intersection of the three events (ABC). A second flood overtopping potential failure mode developed by the risk analysis team could result in overtopping and breach due to misoperation of the pumps without the occurrence of a flood. The potential failure modes of flood overtopping (PFM1) and misoperation (PFM2) are themselves events, whose occurrence or nonoccurrence can be used to describe the state of the system. Assuming no other potential failure modes are plausible, the normal state of the system could be described by the intersection event PFM1' ∩ PFM2' (neither failure mode occurs). Recall that PFM1' is the complement meaning that the failure mode does not occur. The intersection events PFM1 ∩ PFM2' (meaning only PFM1 occurs), PFM1' ∩ PFM2 (meaning only PFM2 occurs), and PFM1 ∩ PFM2 (meaning both PFMs occur) represent three possible failure states for the dam.

A-1.4 Venn Diagrams

The basic concepts of set theory can be illustrated using Venn diagrams. A sample space is typically represented by a rectangle. Events and their relationships are normally depicted on the Venn diagram by overlapping circles or other closed shapes within the sample space. The Venn diagrams in

Figure A-1-1 summarize some basic set theory concepts and operations. Venn diagrams can be developed for a risk analysis to obtain a better depiction and understanding of the relationship between events to support constructing event trees, estimating probabilities, or combining and portraying risks. For example, the relationships between or among multiple potential failure modes can be illustrated using Venn diagrams and described using set theory.

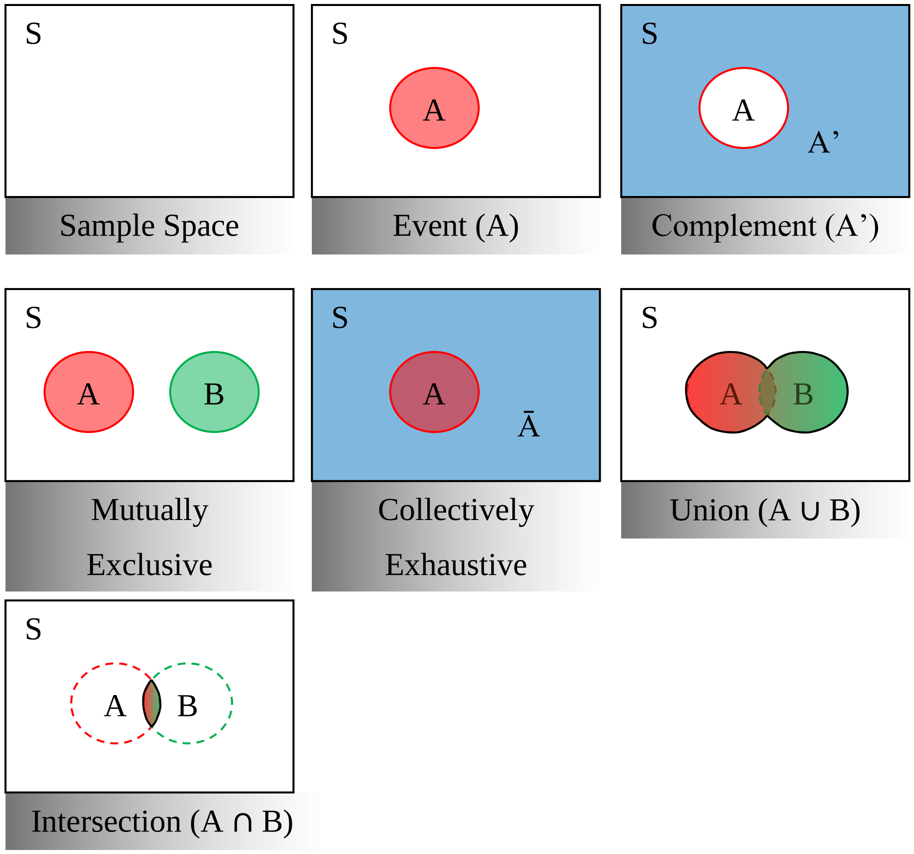

Figure A-1-1 Set Theory Concepts and Operations

A-1.5 Combinatorics

Combinatorics is a branch of mathematics that includes the study of the enumeration, combination, and permutation of set elements. Risk analysis can utilize combinatorics to identify relevant outcomes from a set of possible events.

A-1.5.1 Permutation with Repetition

If each event outcome can be realized more than once and the order of the events does not matter, then the number of permutations is n^k where n is the number of event outcomes available to choose from and k is the number of events that occur. In a river system with two dams and one levee, there are eight permutations for performance of the system. The number of outcomes for each facility is n=2 {breach, no breach} and the number of facilities is k=3 {dam1, dam2, levee}.

The same eight permutations can be also obtained using the binomial coefficient equation below. The equation can be solved in Microsoft Excel using the formula =COMBIN(i,j). Pascal's triangle can also be used as a graphical solution of the binomial coefficient.

$$\binom{i}{j} = \frac{i!}{j!\,(i-j)!}$$ **Equation A-1-1**

where *i!* is the factorial of *i*, 1x2x3…xi. Given the three potential failure modes (j=3), there is 1 combination of zero (i=0) failures, 3 combinations of one (i=1) failure, 3 combinations of two (i=2) failures, and 1 combination of three (i=3) failures. The eight permutations for this example are summarized in Table A-1-1. Once the possible events are enumerated, the analyst can evaluate and decide which of these scenarios might be important and relevant for a risk analysis.

A similar evaluation can be made for an individual dam or levee to enumerate combinations when there is more than one potential failure mode. The number of outcomes for an individual dam or levee is still n=2 {breach, no breach}. The number of events (k) is equal to the number of potential failure modes. An individual levee with three potential failure modes would have eight combinations.

Table A-1-1 Example of Permutation with Repetition

Permutation	Performance for Dam 1	Performance for Dam 2	Performance for Levee
1	No Breach	No Breach	No Breach
2	Breach	No Breach	No Breach
3	No Breach	Breach	No Breach
4	No Breach	No Breach	Breach
5	Breach	Breach	No Breach
6	Breach	No Breach	Breach
7	No Breach	Breach	Breach
8	Breach	Breach	Breach

A-1.5.2 Permutation Without Repetition

If each event outcome can be realized only once and the order of the events does matter, then the number of permutations within a subset of m events is $\frac{i!}{(i-m)!}$ where i is the total number of events. For example, permutation number five from the previous example in Table A-1-1 could have six additional permutations (i=m=3) if the order of the events is important. Perhaps the consequences are different depending on which dam breach occurs first and whether or not the levee overtops before or after a dam breaches. The six additional permutations are listed in Table A-1-2. Once the possible events are enumerated, the risk analyst can evaluate and decide which of these scenarios might be important and relevant for a risk analysis. Perhaps the order in which the breaches occur affects consequences but the order of non-breach events does not. This might lead to a conclusion that permutations 5-1 and 5-3 are relevant for the risk analysis and the rest can be screened out.

Table A-1-2 Permutation Without Repetition Example

Permutation	First Event	Second Event	Third Event
5-1	Dam 1	Dam 2	Levee’
5-2	Dam 1	Levee’	Dam 2
5-3	Dam 2	Dam 1	Levee’
5-4	Dam 2	Levee’	Dam 1
5-5	Levee’	Dam 1	Dam 2
5-6	Levee’	Dam 2	Dam 1

Levee’ means the levee does not breach

A-1.6 Probability

A-1.6.1 Axioms

Probability theory is founded on three axioms that have been attributed to Andrei Kolmogorov. The first axiom states that the probability of an event (A) is a non-negative real number. The second axiom states that the probability of an event that is certain to occur is equal to one. The third axiom (sometimes referred to as the addition rule) states that the union of two or more mutually exclusive events is equal to the sum of the probabilities for each event. The axioms are summarized by the three equations below.

$$\text{First Axiom:} \quad P(A) \geq 0$$

$$\text{Second Axiom:} \quad P(S) = 1$$

$$\text{Third Axiom:} \quad P(A \cup B) = P(A) + P(B)$$

The third axiom can be expressed as a multiplication rule instead of the addition rule. In practice, it makes no difference because all of the remaining probability formulas can be derived from either set of axioms. The multiplication rule can be expressed using the equation below.

$$\text{Multiplication Rule:} \quad P(A \cap B) = P(A)\, P(B|A)$$

In the above equation, $P(B|A)$ is the conditional probability of event B given that event A occurs. In a risk analysis, this might correspond to the probability that a levee will breach given that a 50 year flood occurs, expressed as P(Breach | 50 Year Flood). Note that the Annualized Failure Probability (AFP) is calculated using the multiplication rule as the intersection probability of the events that comprise a potential failure mode.

A-1.6.2 Interpretation

Frequency and degree-of-belief describe two broad interpretations of probability. In practice, probability estimates used in risk analysis are based on degree-of-belief. Frequency based probabilities can be used to inform these degree-of-belief probability estimates. Both interpretations of probability are useful in risk analysis and both follow the same probability calculus.

Frequency probabilities are based on a stable frequency for the occurrence of an event over a long sequence of trials. Frequency probabilities can also be estimated based on the physical properties of a system. The frequency of rolling a 2 using a 6 sided die should be 1/6. This can be estimated directly by rolling the die many times or indirectly by concluding the die is fair based on its physical properties. Similarly, observations can be used to estimate probabilities for a dam or levee risk analysis. If damage to clay tile drains has been observed at 50 out of 100 dams inspected, then a risk analyst might estimate the probability of clay tile drain damage at the dam under consideration to be about 50% as long as the conditions at the dam being evaluated are reasonably consistent with the conditions at the inspected dams. Without a physical basis like the fair die example, a sufficient number of observations is required to obtain a reasonably accurate probability estimate.

Degree-of-belief probabilities are based on a rational weighting of evidence that can be manifested by a willingness to take a particular action, to bet at particular odds, or to consider particular odds as fair. Personal experience, expert judgment, and other manifestations of deductive and inductive reasoning can be used as a basis for estimating degree-of-belief probabilities. For example, a degree-of-belief interpretation might arrive at the same probability of 1/6 for rolling a 2 based on the available evidence (assumption that the die is fair, visual observation of the die characteristics, measurement of the die properties, or past experience with similar looking die). Similarly, an expert might combine their general knowledge of the internal erosion mechanism (physics of the process, more likely and less likely factors, knowledge of past incidents) with the specific characteristics of an embankment dam (construction practices used, soil properties, location of the phreatic surface) to estimate the probability of internal erosion initiation under a particular loading.

A-1.6.3 Uncertainty

Two general types of uncertainty can be described as aleatory (natural variability or randomness) and epistemic (knowledge uncertainty). Aleatory uncertainty characterizes processes that are assumed to be random in time or space. The occurrence of floods might be assumed to be random in time and the spacing of joints in a bedrock foundation might be assumed to be random in space. In practice, aleatory uncertainty is treated as irreducible. In other words, there is no practical way to reduce the uncertainty through the acquisition of more knowledge. Epistemic uncertainty characterizes our lack of knowledge regarding the state of nature. A possible foundation flaw either exists or does not exist, but we don't have sufficient knowledge to determine for certain whether or not the flaw exists. Epistemic uncertainty considers the uncertainty in both models and model parameters. Uncertainty in modeling includes our ability to identify a proper model, the ability of the model to represent reality, and our understanding of how the model may be changing over time. Uncertainty in model parameters includes our ability to identify the appropriate representative parameters and consistently estimate values for the parameters through observation or measurement. In practice, epistemic uncertainty is treated as reducible. In other words, more knowledge can be obtained to reduce the magnitude of the uncertainty. Additional exploration could reduce uncertainty in the possible presence of a foundation flaw. These uncertainty concepts are applicable not only to risk analysis models and risk estimates but also to decision making processes.

A-1.6.4 Expressing Probability

Probability estimates can be expressed as a percent (10% chance), as a fraction (1/10 chance), as a decimal (0.1 probability), or as odds (1:9). Each of these four values has the same probability and the same meaning. Probabilities that apply to an annual time period can be expressed as an annual exceedance probability (AEP). This is common for the characterization of flood and seismic hazards. Probabilities can also be defined as a function of time to describe temporal processes such as climate change or degradation (e.g. corrosion).

A-1.6.5 Random Variables

A random (or stochastic) variable is used to represent an uncertain quantity whose value can take on a number of possible values. The uncertainty associated with the random variable could be the result of natural variability or a lack of knowledge. Despite the name, random variables do not necessarily have to describe a random process. For example, the magnitude of a spring flood might be modeled as a random process that varies from year to year, whereas a fault in the dam foundation might be modeled as a lack of knowledge. We do not know whether or not it exists. Both of these scenarios can be described using random variables. The flood might be described by a range of peak discharge values and the presence of the fault might be described by two scenarios, either ‘yes, it exists’ or ‘no, it does not exist’.

A-1.6.6 Combining Probabilities (Union of Two or more Events)

The probability for the union of two (or more) events is a common calculation in risk analysis. It is used to sum probabilities and risks across multiple hazards, event tree branches, and potential failure modes.

A general equation for calculating the probability of the union of two events is shown below. The summation of the three terms contained within the {} brackets represent the probabilities for the occurrence of event A only, the occurrence of event B only, and the occurrence of events A and B. The equation can be expanded for three or more events. The number of terms needed in the equation is equal to 2^n-1, where n is the number of events. The calculation becomes more cumbersome and complex as the number of events increases.

$$P(A \cup B) = \{P(A) - P(A \cap B)\} + \{P(B) - P(A \cap B)\} + \{P(A \cap B)\}$$

For only two events, a simplified form of the same equation is more commonly found in the literature.

$$P(A \cup B) = P(A) + P(B) - P(A)P(B)$$

Two (or more) events are mutually exclusive when both events cannot occur in the same experiment or trial. The equation below for the probability of the union of two mutually exclusive events is the same as the third probability axiom.

$$P(A \cup B) = P(A) + P(B)$$

This equation can also be obtained from the general union equation by recognizing that $\mathrm{P}(A \cap B) = 0$ for mutually exclusive events. If the events A and B cannot both occur, then the probability of the event AB must equal zero.

Three ways the union calculation is applied in a risk analysis are as follows. First, event tree branches are, by definition, mutually exclusive events. This allows one to simply sum probabilities and risks across branches to obtain a total risk estimate for all loadings and PFMs included in the tree. Second, floods and earthquakes are typically modeled as mutually exclusive events in a risk analysis. This allows estimating the risks separately for each hazard and then summing them to obtain a total risk estimate. This simplifying assumption is not strictly true, and may not be appropriate or valid in every situation, but it is typically reasonable to assume that the probability of a major earthquake occurring coincident with a major flood is negligible. Third, potential failure modes resulting from a particular loading are often modeled as mutually exclusive events in a risk analysis. This provides a basis for estimating risks for individual PFMs in separate event trees and then summing the estimates to obtain a total risk estimate. Strictly speaking, hazards and PFMs are not mutually exclusive. This is merely a simplifying assumption that can be made when the joint probability [i.e. P(A)P(B)] and associated consequences of the intersection event is relatively small [i.e. P(A)P(B) << $P(A \cup B)$] such that the intersection event can be omitted from the risk analysis. When events cannot be reasonably modeled as mutually exclusive, the general union equation can be used.

A-1.6.7 Combining Probabilities (Intersection of Two or more Events)

The probability for the intersection of two (or more) events is a common calculation in risk analysis. It represents the probability that both events occur and is used to multiply probabilities to obtain the probability for the outcome of a sequence or collection of events. This is the basis for multiplying probabilities along event tree pathways and for multiplying probabilities for the sequence of events associated with a potential failure mode. For example, a breach can occur only if all of the underlying events that define the potential failure mode occur. The probability of breach can be calculated as the product of the probabilities for the underlying events that define the PFM.

A general equation for calculating the probability of the intersection of two events is shown below. This is the third probability axiom expressed as the multiplication rule. The two terms represent the probability that event A occurs multiplied by the probability that event B occurs given that event A occurs. The equation can be expanded for three or more events. The number of terms needed in the equation is generally equal to the number of events.

$$P(A \cap B) = P(A)P(B|A)$$

Two (or more) events are statistically dependent if the occurrence of one event affects the occurrence probability of the other event(s). This is why the second term in the general equation is P(B|A). This is the probability that B occurs given that A occurs. In a typical event tree, this might be the probability that internal erosion initiates given a particular flood loading occurs. The intersection probability for the flood loading and initiation would be the probability of the flood load multiplied by the probability of initiation given (i.e. conditional on) the flood load. The probability of initiation depends on the magnitude of the flood load.

Two (or more) events are statistically independent when the occurrence of one event does not affect the probability for the other event(s). When events are independent, P(B|A) = P(B). Event A does not influence the probability of event B occurring. Potential failure modes are typically estimated assuming they are statistically independent. This is an assumption that simplifies the estimation of risk. Strictly speaking, this may not always be a valid simplifying assumption. For example, a spillway erosion potential failure model might reduce the likelihood of an overtopping failure mode. If spillway erosion occurs, the outflow might increase making overtopping less likely.

A-1.6.8 Combining Probabilities (DeMorgan's Rule)

A common formula used to calculate a total probability is derived from DeMorgan's rule. Examples include calculating the total probability of failure given multiple PFMs or calculating the probability of flooding over a 30-year mortgage period. For two events A and B, DeMorgan's rule states that the complement of the union of the two events is equal to the intersection of their complements. This can be expressed by the equation below.

$$(A \cup B)' = A' \cap B'$$

The Venn diagrams shown in Figure A-1-2 provide a conceptual derivation of DeMorgan's rule.

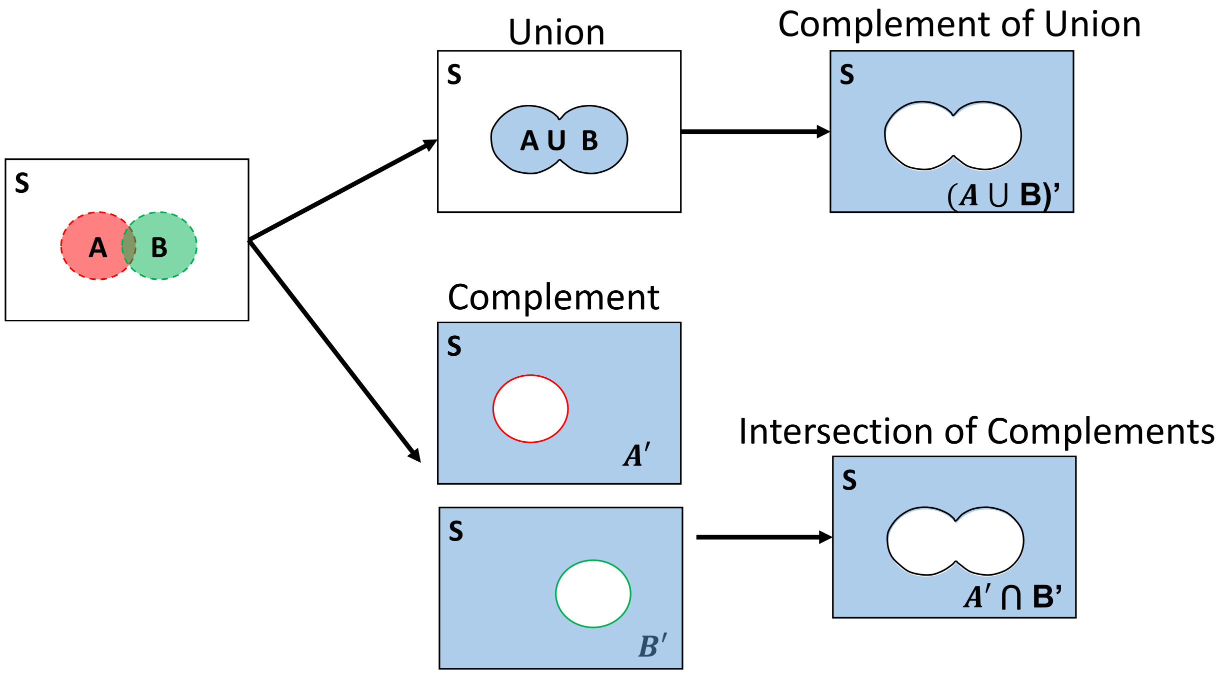

Figure A-1-2 DeMorgan's Rule

In practice, DeMorgan's rule can be applied to simplify some risk analysis calculations. For example, the total probability of failure for a system with n PFMs can be calculated using the equation shown below. This can simplify the calculation of total risk when the number of potential failure modes is greater than two. Recall that the union equation becomes cumbersome when there are more than two events (in this case more than two PFMs). A conceptual derivation of the total probability equation using DeMorgan's rule is shown in Figure A-1-3.

$$P(System\ Failure) = 1 - P(No\ System\ Failure) = 1 - \prod_{i=1}^{n} [1 - P(PFM_i)]$$

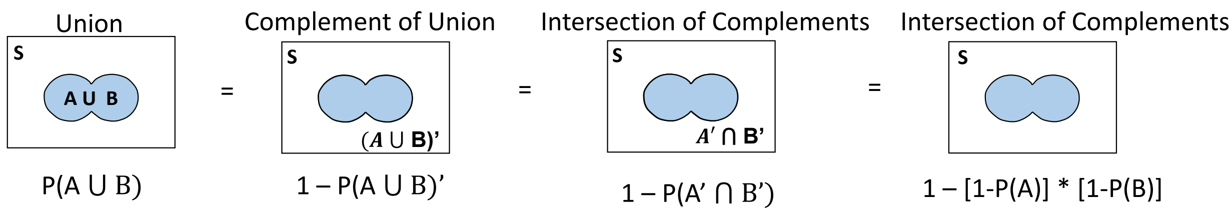

Figure A-1-3 Application of DeMorgan's Rule

A-1.6.9 Combining Probabilities (Uni-Modal Bounds)

The unimodal bounds theorem (Ang and Tang, 1984) states that for 'n' positively correlated events ($E_1, E_2, E_3, \ldots, E_n$) with corresponding probabilities [$P(E_1), P(E_2), P(E_3), \ldots, P(E_n)$], the total probability for the union of the events [$P(E) = P(E_1 \cup E_2 \cup E_3 \ldots \cup E_n)$] lies between the upper and lower bounds given by the following equation.

$$\max\,[P(E1), P(E2), P(E3), \ldots, P(En)] \leq P(E) \leq 1 - \prod_{i=1}^{n}[1 - P(E_i)]$$

The upper bound on the right side of the equation is based on a calculation of the total probability of system failure using DeMorgan's rule. The lower bound on the left side of the equation is based on the individual event with the largest probability. Events that are correlated will yield a total probability closer to the lower bound. Events that are uncorrelated will yield a total probability that is closer to the upper bound. In practice, the degree of correlation can be difficult to estimate. It is common for risk analysts to assume the upper bound value. This assumption may not be appropriate in all situations.

A-1.6.10 Combining Probabilities (Central Limit Theorem)

When statistically independent random variables are summed, the distribution of the sum will trend toward a normal distribution even if the distributions of the variables being summed are not normal. When statistically independent random variable are multiplied, the distribution of the product will trend toward a log-normal distribution even if the distributions of the variables being multiplied are not normal. This is why the distribution of C-D (capacity minus demand) in a reliability analysis will typically trend toward a normal distribution. The annual probability of failure for a potential failure mode will also typically trend toward a normal distribution because the end branches of an event tree are summed to obtain the total. The distribution for factor a safety (C/D) will typically trend toward a log-normal distribution because capacity is divided by demand. The end branches of an event tree will also typically trend toward a log-normal distribution because probabilities along an event tree pathway are multiplied.

A-1.7 Statistics

A-1.7.1 Probability Distributions

The event described by a particular value (or range of values) of a random variable must be expressed with an associated probability. Probability distributions can be used to describe the probabilities associated with the possible values of a random variable. For example, we can estimate a probability distribution for the annual maximum ground acceleration at a dam site, the permeability of a sand layer in a levee foundation, the system response for a potential failure mode, or the effectiveness of an evacuation warning. Virtually all parameters considered and applied in a risk analysis have some degree of uncertainty and are therefore candidates to consider modeling them as random variables.

Random variables can be discrete or continuous. Discrete random variable can have a finite number of possible values (e.g. number of monoliths that breach). Continuous random variables can have an infinite number of possible values (e.g. peak ground acceleration at a levee). For discrete random variables, a probability of occurrence can be estimated and assigned to each of the possible outcomes. The number of spillway gates that fail to open on demand could be modeled as a discrete random variable. A probability mass function (PMF) is commonly used to describe the probability distribution for a discrete random variable. A possible probability mass function for the spillway gates is shown in Figure A-1-4. The probability that exactly one gate does not open can be obtained directly from the probability mass function as about 0.35. The probability that zero or one gate does not operate can be obtained by summing the probability mass for zero and one gate which would be about 0.53+0.35=0.88. The probabilities for all of the gate scenarios must sum to 1.0 to satisfy the probability axioms.

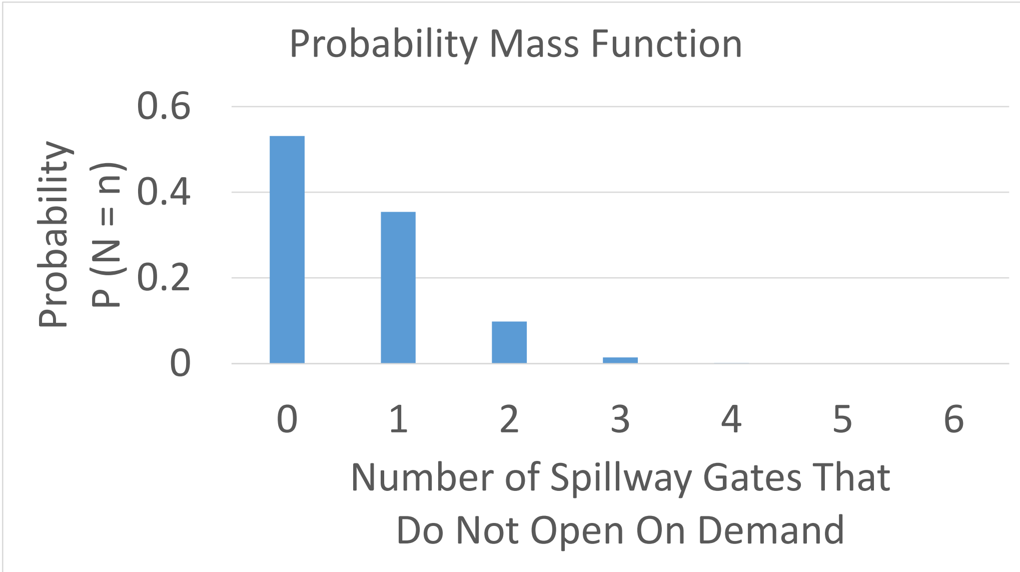

Figure A-1-4 Probability Mass Function for a Discrete Random Variable

A cumulative distribution function (CDF) is another way to describe the same probability distribution for the spillway gates. The cumulative distribution describes the probability that the random variable is less than or equal to a particular value. The cumulative distribution function for the spillway gates is shown in Figure A-1-5. The probability that exactly one gate does not open can be obtained from the cumulative distribution function. The probability of 0.88 for one gate or less includes the events {0 gate, 1 gate}. The probability of 0.53 for zero gates includes the event {0 gate}. The difference between these two probabilities is the event for exactly one spillway gate {1} which has a probability of about 0.88-0.53=0.35. This is the same value that was estimated previously from the probability mass function. The probability that zero or one gate does not operate, which is equivalent to less than or equal to one gate on the cumulative distribution, can be obtained directly as about 0.88. The probability of six or less gates must equal one to satisfy the probability axioms.

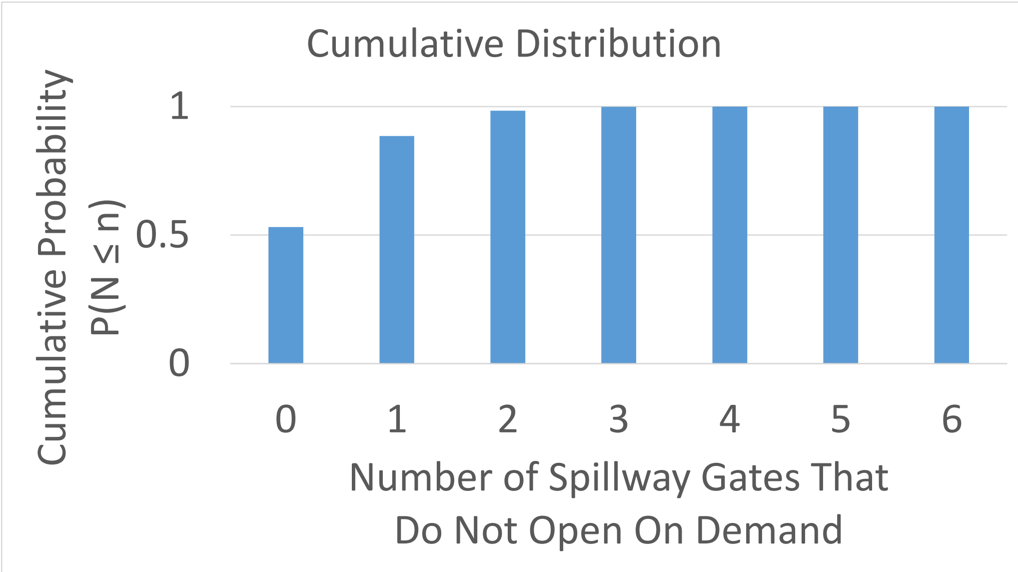

Figure A-1-5 Cumulative Distribution Function for a Discrete Random Variable

Continuous random variables can take on an infinite number of possible values. For example, the thickness of a sand layer in a levee foundation might take on any value greater than or equal to zero. A probability density function (PDF) can be used to describe the probability distribution for continuous random variables. A possible probability density function for sand layer thickness is shown in Figure A-1-6.

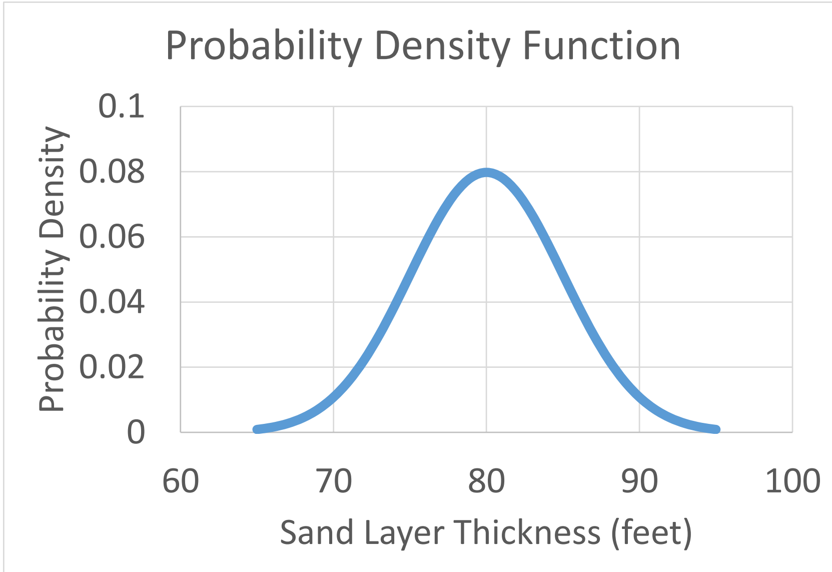

Figure A-1-6 Probability Density Function for Continuous Random Variable

The probability that the thickness will fall between two particular values can be obtained from the probability density function as the integral (or area) between the two values. This is shown graphically in Figure A-1-7 and expressed mathematically with the following equation.

Equation A-1-2

$$P(a < X \leq b) = \int_a^b f_X(x)dx$$

In the example, the probability that the sand layer thickness is between 75 and 85 feet is equal to 0.68. The probability for any specific value (say a thickness of exactly 80 feet) is equal to zero for any continuous random variable. The units on the vertical axis are a density (not a probability). This is why continuous random variables, such as those typically used to characterize flood or seismic hazard, must be evaluated in a risk analysis using partitions (also commonly referred to as load ranges or load intervals). The total area under a probability density function must equal one to satisfy the probability axioms.

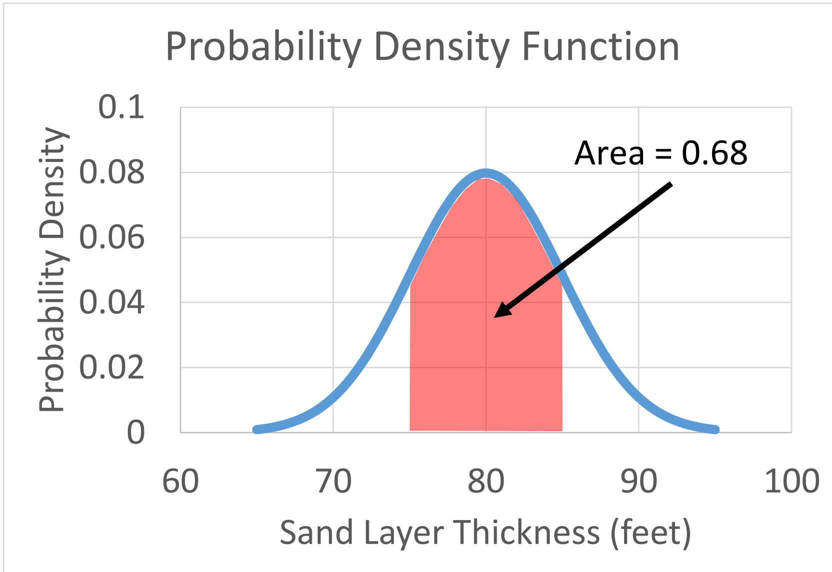

Figure A-1-7 Probability Density Function for a Continuous Random Variable

Continuous random variables can also be described by a cumulative distribution function. The cumulative distribution describes the probability that the random variable is less than or equal to a particular value. This is the integral of the probability density function over all values less than or equal to the value of interest. The integral can be expressed mathematically with the following equation.

$$P(-\infty < X \leq b) = \int_{-\infty}^{b} f_X(x)dx$$

Equation A-1-3

A possible cumulative distribution function describing the system response for a potential failure mode at a dam is shown in Figure A-1-8. The probability that the dam will breach when subjected to a peak water load of 1500 feet is equal to the probability that the capacity (or strength) of the dam is less than or equal to the demand (or load). In this example, the probability of failure is 0.16. The cumulative probability function must have an upper bound of one to satisfy the probability axioms.

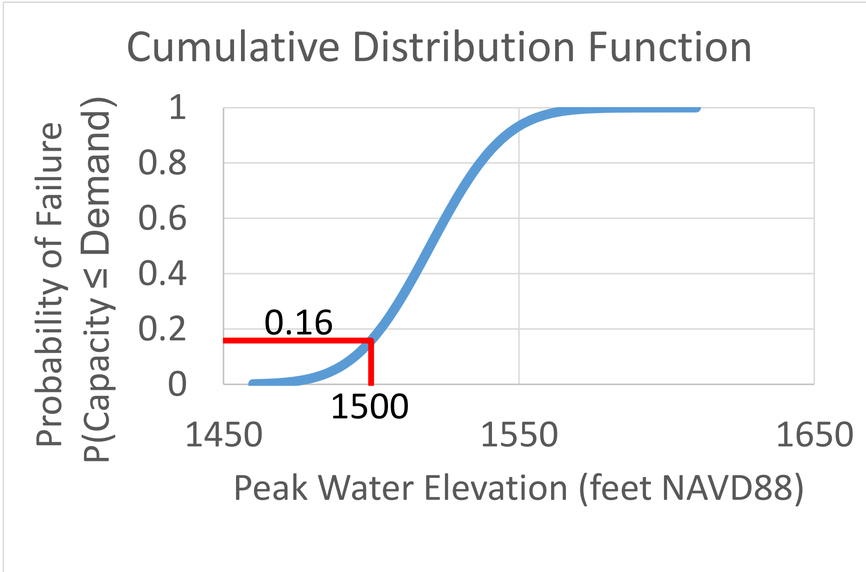

Figure A-1-8 Cumulative Distribution Function for a Continuous Random Variable

A survivor function (also called a complementary cumulative distribution function or exceedance curve) can be used to describe the probability that a particular value for the random variable will be greater than a particular value. For this reason, survivor functions are sometimes referred to as exceedance functions. Flood and seismic hazards are typically defined this way. The survivor function can be expressed mathematically as an integral of the probability density function with the following equation.

Equation A-1-4

$$P(-\infty < X \leq b) = \int_{-\infty}^{b} f_X(x)dx$$

Greater than is the customary sign convention for survivor functions in the United States with one important exception. The F-N chart, which is a survivor function for life loss, is defined as the annual probability of life loss greater than or equal to a particular life loss value.

The survivor function can also be derived from the cumulative distribution function using the following equation by recognizing that the survivor function is the complement of the cumulative distribution function. For this reason, survivor functions are sometimes referred to as complementary cumulative distribution functions.

$$P(a < X \leq \infty) = 1 - P(-\infty < X \leq a)$$ **Equation A-1-5**

A possible survivor function for the seismic hazard at a dam is shown in Figure A-1-9. In this example, the probability that the maximum ground acceleration in a given year will be greater than 0.2g is equal to 4.6E-3 (about 1 in 220). Recall that the probability of an acceleration exactly equal to 0.2g is zero. The probability of an acceleration between 0.1g and 0.2g in a given year can be calculated from the survivor function using a loading partition. In this example, the probability for a maximum acceleration between 0.1g and 0.2g would be calculated as the difference between the AEP for 0.1g and the AEP for 0.2g. (2.3E-2 – 4.6E-3 = 0.018). Strictly speaking, an AEP describes the probability of one or more events occurring in a given year. However, the probability of more than one event occurring is typically negligible for AEPs less than about 0.1 (greater than a 10 year event). Alternative techniques such as partial duration series are available when the AEP is greater than 0.1 and the possible occurrence of more than one event in a given year is important for characterizing the risk.

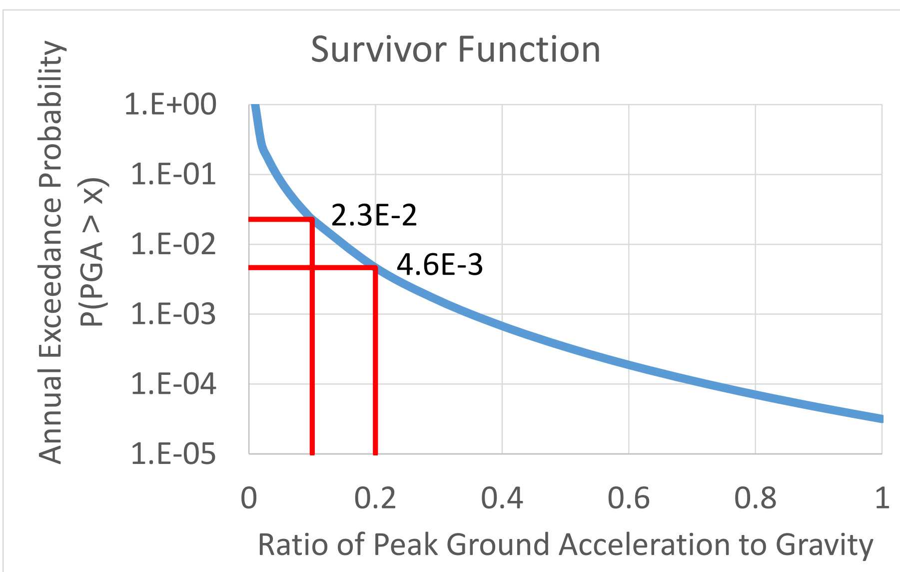

Figure A-1-9 Survivor Function for a Continuous Random Variable

A-1.7.2 Moments

A moment is a quantitative measure of shape used in both statistics and mechanics. Moments can be estimated from observations, measurements, or expert opinion. The first moment is the mean (or centroid) which measures the central tendency of data or a distribution. The second moment is the variance (or rotational inertia) which measures the amount of spread about the mean. The third moment is the skewness which measures the asymmetry about the mean. Equations for estimating these moments are provided in Table A-1-3.

Table A-1-3 Moment Equations

Moment	Meaning	Discrete Random Variable	Continuous Random Variable
Mean (First)	Center of Mass	$\bar{x} = \sum_{1}^{n} x_i p(x_i)$	$\bar{x} = \int_{-\infty}^{\infty} x f(x) dx$
Variance (Second)	Central tendency	$\sigma^2 = \sum_{1}^{n} (x_i - \bar{x})^2 p(x_i)$	$\sigma^2 = \int_{-\infty}^{\infty} (x - \bar{x})^2 f(x) dx$
Skew (Third)	Symmetry	$\gamma = \frac{\sum_{1}^{n} (x_i - \bar{x})^3 p(x_i)}{\sigma^3}$	$\gamma = \frac{\int_{-\infty}^{\infty} (x - \bar{x})^3 f(x) dx}{\sigma^3}$

Other common statistical measures include the median, mode, and geometric mean. The median is the 50th percentile of a distribution which means that there is equal probability of a value greater than or less than the median. The mode is the most probable value which means that the mode has the largest probability mass (discrete) or probability density (continuous). A graphic depiction of the mean, median, and mode is shown in Figure A-1-10.

Additional parameters can be obtained from the set of basic parameters. The standard deviation is equal to the square root of the variance. The coefficient of variation is equal to the standard deviation divided by the mean.

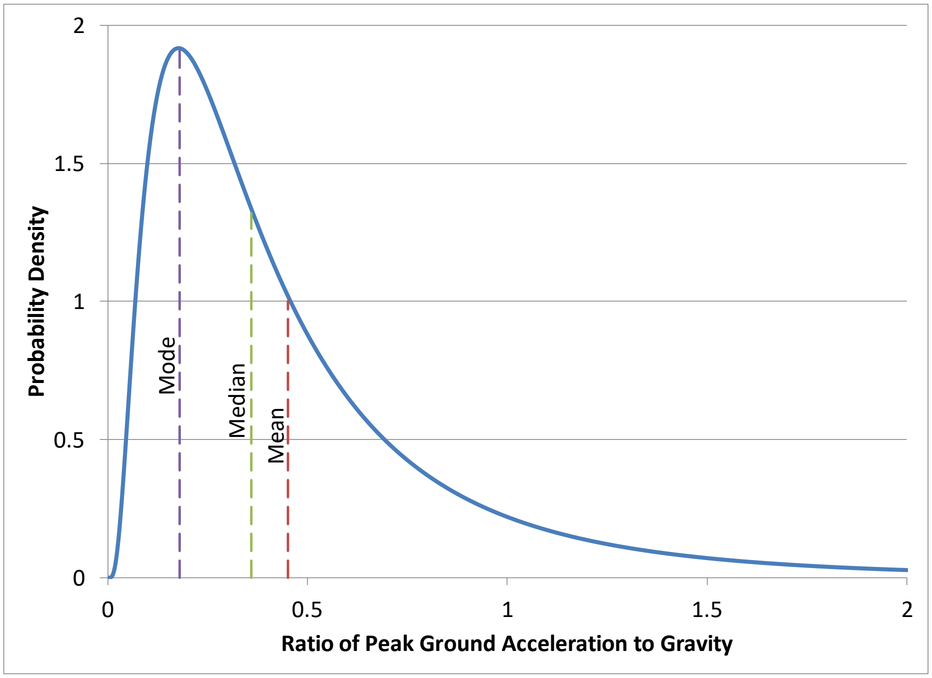

Figure A-1-10 Mean, Median, and Mode

The geometric mean is defined as the *n*th root of the product of *n* data values as shown in the equation below. It is typically used to calculate an average value in log space because the geometric mean is equal to the exponential of the arithmetic mean of the logarithms.

Equation A-1-6

$$G = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = e^{\left(\frac{\sum_{i=1}^{n} \ln x}{n}\right)}$$

For example, the mean estimate for an SQRA probability estimate between 1E-4 and 1E-5 should be calculated as a geometric mean using the equation below. The geometric mean is also commonly used to calculate the average elevation or acceleration for a load partition because flood and seismic hazard curves typically have a logarithmic form.

$$G = (10^{-4} * 10^{-5})^{\frac{1}{2}} = 3x10^{-5}$$

A-1.7.3 Analytical Distributions

A multitude of probability distribution are available to describe random variables. Some of the more common types which might be useful in a risk analysis include uniform, triangular, normal, log-normal, PERT, and Weibull. Other distributions are available and may be more appropriate for a particular application. When selecting an analytical distribution, the following list of questions can be used as a guide.

- What distribution provides the best fit?
- What distribution should be expected?
- What distributions have fit well in similar situations?
- What do the experts think?
- Does it matter?

The uniform distribution (or rectangular distribution) is a continuous two parameter distribution that can be used to describe a range of values that are equally probable. The distribution is defined by a lower bound (a) and an upper bound (c). The probability density function has a constant value of $\frac{1}{c-a}$ for any value between a and c. The cumulative distribution function has a value of $\frac{x-a}{c-a}$ for any value x that is between a and c. A uniform distribution is shown in Figure A-1-11.

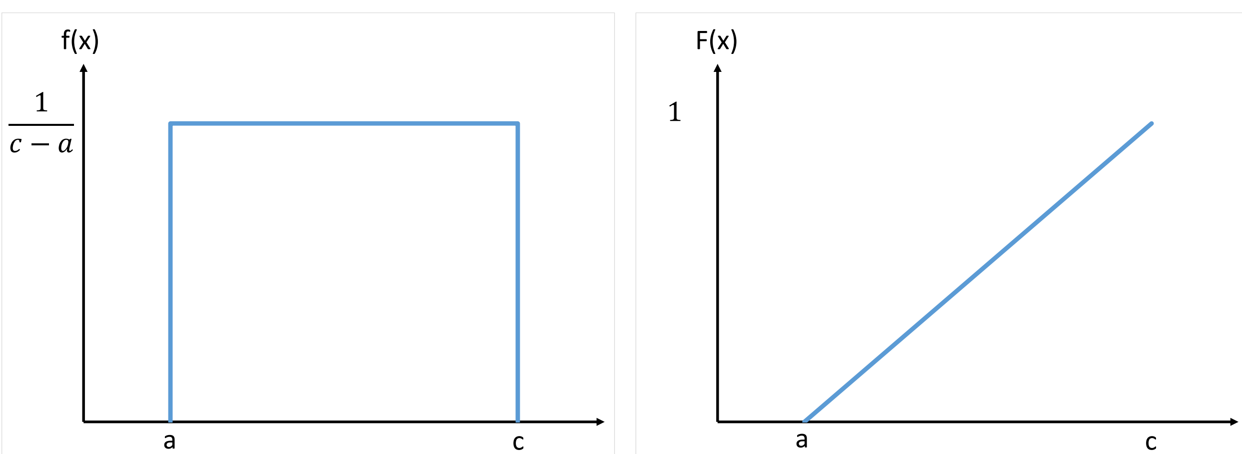

Figure A-1-11 Uniform Distribution

The triangular distribution is a continuous three parameter distribution defined by a lower bound (a), a most likely (mode) value (b), and an upper bound (c). The probability density function has a value of $\frac{2(x-a)}{(c-a)(b-a)}$ for a≤x≤b and a value of $\frac{2(c-x)}{(c-a)(c-b)}$ for b≤x≤c. The cumulative distribution has a value of $\frac{(x-a)^2}{(c-a)(b-a)}$ for a≤x≤b and a value of $1-\frac{(c-x)^2}{(c-a)(c-b)}$ for b≤x≤c. A triangular distribution is shown in Figure A-1-12.

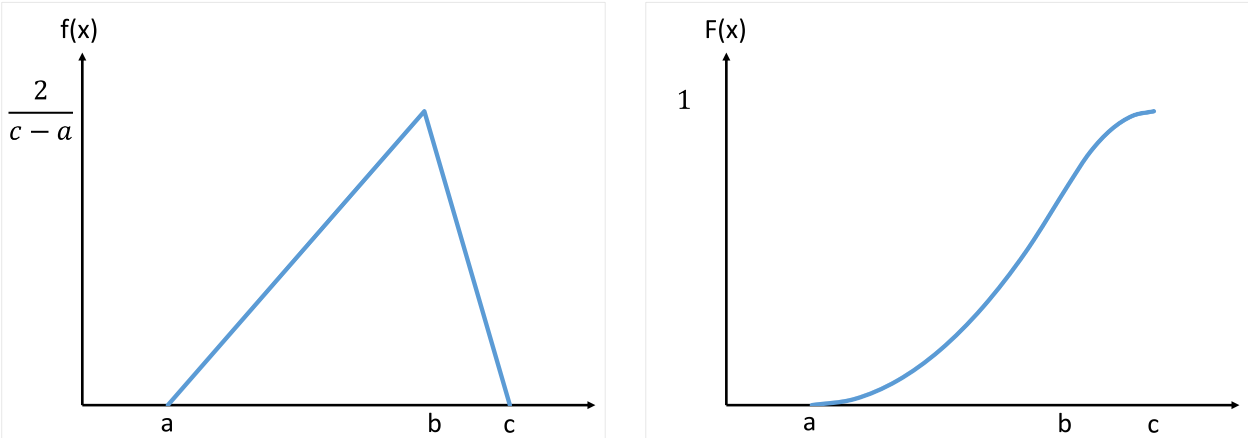

Figure A-1-12 Triangular Distribution

The normal distribution is a two parameter continuous distribution defined by a mean or expected value (μ) and a variance (σ^2) or standard deviation (σ). The random variable can take on any value between -∞ and +∞. Physical quantities that result from a summation of many independent processes have distributions that are approximately normal. The probability density function for the normal distribution is given by the equation below. The probability density can be calculated with Microsoft Excel using the formula =NORM.DIST(x,μ,σ,FALSE).

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$ **Equation A-1-7**

The cumulative distribution function for the normal distribution is given by the equation below. The cumulative distribution can be calculated with Microsoft Excel using the formula =NORM.DIST(x,μ,σ,TRUE). A normal distribution is shown in Figure A-1-13.

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{\frac{-1}{2v^2}} d$$

Equation A-1-8

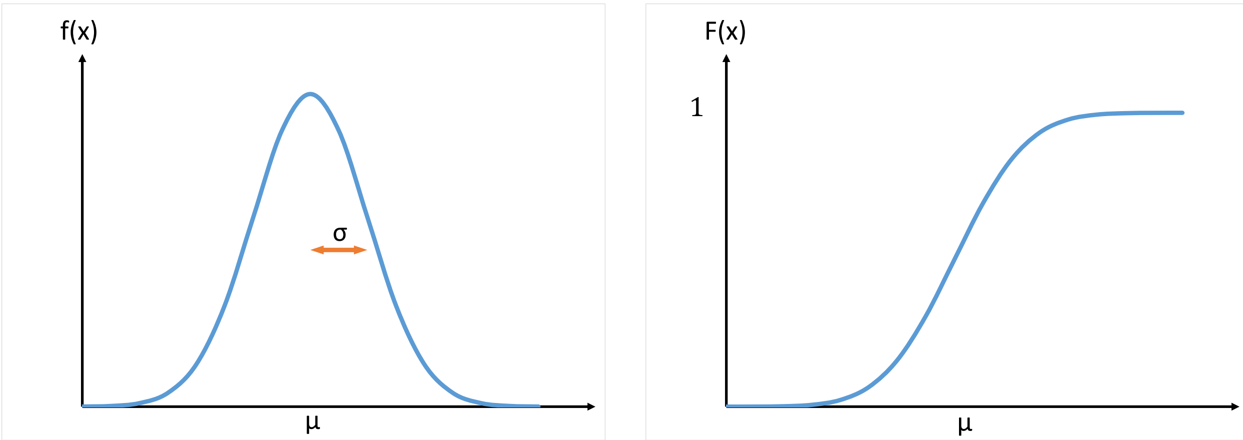

Figure A-1-13 Normal Distribution

The log-normal distribution is a two parameter continuous distribution for a random variable whose logarithm has a normal distribution. The parameters μ and σ are the mean and standard deviation for ln(x). The relationships between these parameters and the mean (m) and standard deviation (s) of x is given by the equations below.

$$\mu = ln\left(\frac{m}{\sqrt{1 + \frac{s^2}{m^2}}}\right)$$

Equation A-1-9

$$\sigma = \sqrt{ln\left(1 + \frac{s^2}{m^2}\right)}$$

Equation A-1-10

The random variable can take on any value between 0 and +∞. Physical quantities that result from a product of many independent processes typically have distributions that are approximately log-normal. The probability density function for the log-normal distribution is given by the equation below. The probability density can be calculated with Microsoft Excel using the formula =LOGNORM.DIST(x,μ,σ,FALSE).

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$ **Equation A-1-11**

The cumulative distribution function for the log-normal distribution is given by the equation below. The cumulative distribution can be calculated with Microsoft Excel using the formula =LOGNORM.DIST(x,μ,σ,TRUE). A log normal distribution is shown in Figure A-1-14.

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\ln x - \mu}{\sigma}} e^{\frac{-1}{2v^2}}\, dv$$ **Equation A-1-12**

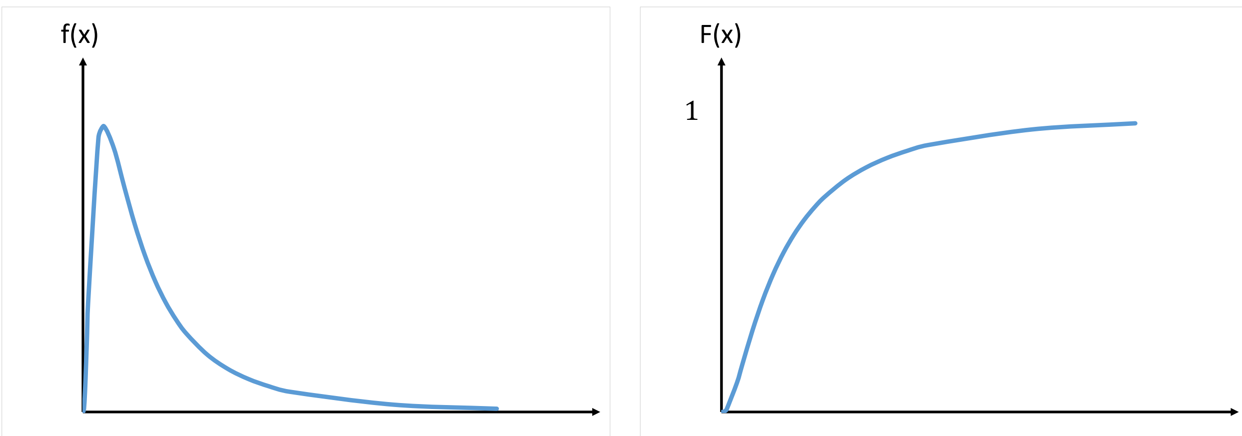

Figure A-1-14 Log-Normal Distribution

The PERT distribution was developed specifically for use in expert elicitation. It is a continuous three parameter distribution defined by a lower bound (a), an upper bound (c), and a most likely (mode) value (b). The PERT distribution has a smoother shape than the triangular distribution and a different mean. It is derived from the four parameter beta distribution with a mean defined by the following equation.

$$\mu = \frac{a + 4b + c}{6}$$ **Equation A-1-13**

The probability density function for the PERT distribution is given by the equation below.

$$f(x) = \frac{(x-a)^{\alpha-1}(c-x)^{\beta-1}}{B(\alpha,\beta)(c-a)^{\alpha+\beta-1}}$$ **Equation A-1-14**

Parameters for the above equation can be calculated using the equations below. The gamma function (Γ) can be calculated with Microsoft Excel using the formula =Gamma()

$$\alpha = \frac{4b + c - 5a}{c - a}$$ **Equation A-1-15**

$$\beta = \frac{5c - a - 4b}{c - a}$$ **Equation A-1-16**

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$ **Equation A-1-17**

The cumulative distribution function for the PERT distribution is given by the equation below.

$$F(x) = I_x(\alpha,\beta)$$ **Equation A-1-18**

The incomplete beta function (I) can be calculated with Microsoft Excel using the formula =BETA.DIST(x,α,β,1,a,c). A PERT distribution is shown in Figure A-1-15.

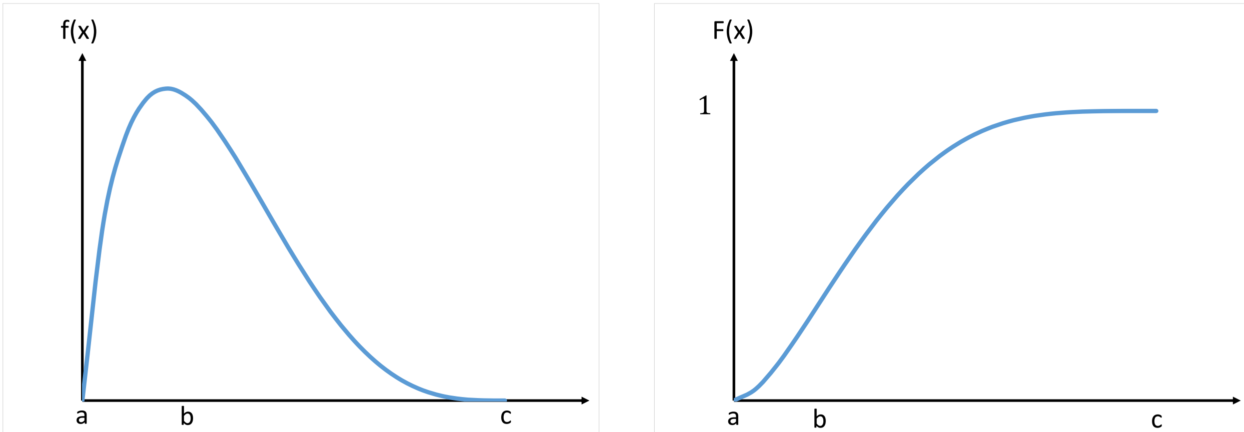

Figure A-1-15 PERT Distribution

Bathtub curves are commonly used in reliability engineering to describe a change in failure rate of electrical or mechanical components over time. The first part of the curve typically has a decreasing failure rate and is sometimes referred to as the early failure period. The second part of the curve typically has a constant failure rate and is sometimes referred to as the random failure period. The third part of the curve typically has an increasing failure rate and is sometimes referred to as the wear out period. A bathtub curve can be constructed from the Weibull distribution which is defined by a shape (β) and a scale(η) parameter. The probability density function for the Weibull distribution is given by the equation below.

$$f(x) = \frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1} e^{-\left(\frac{x}{\eta}\right)^{\beta}}$$ **Equation A-1-19**

The cumulative distribution function for the Weibull distribution is given by the equation below.

$$F(x) = 1 - e^{-\left(\frac{x}{\eta}\right)^{\beta}}$$ **Equation A-1-20**

The failure rate can be calculated using the following equation. The failure rate decreases with time when $\beta<1$. The failure rate is constant when $\beta=1$. The failure rate increases with time when $\beta>1$.

$$\lambda(T) = \frac{\beta}{\eta}\left(\frac{T}{\eta}\right)^{\beta-1}$$ **Equation A-1-21**

The mean time to failure can be calculated using the following equation.

$$\mu_T = \eta\Gamma\left(\frac{1}{\beta} + 1\right)$$ **Equation A-1-22**

A bathtub curve derived from a Weibull distribution is shown in Figure A-1-16.

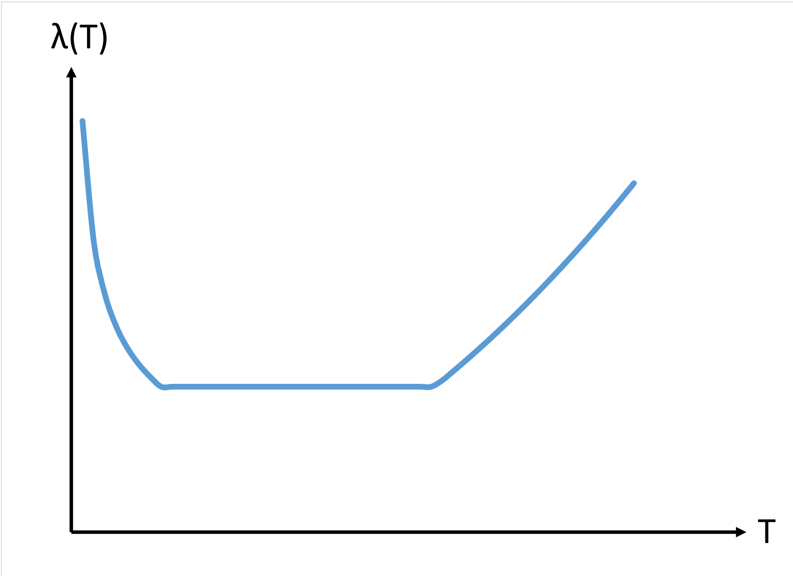

Figure A-1-16 Failure Rate (Bathtub Curve)

A-1.7.4 Confidence Limits and Intervals

The uncertainty associated with an estimated or sampled value of a random variable can be described using confidence limits and confidence intervals. The confidence interval [a,b] for a specified degree of confidence (C%) can be calculated based on the equation below.

$$C\% = \left(\int_a^b f(x)dx\right) * 100$$ **Equation A-1-23**

A confidence limit can be obtained by replacing a in the equation above with the lower bound of the distribution. Confidence limits less than 50% are referred to as lower confidence limits. Confidence limits greater than 50% are referred to as upper confidence limits. Confidence limits and intervals are typically reported based on equal tail probabilities. For example, a confidence interval of 90% would typically be calculated based on an upper confidence limit of 95% and a lower confidence limit of 5%.

For example, assume the friction angle for the soils at a particular site are sampled. Based on the sample data, the friction angle is believed to fit a normal distribution with a mean of 32° and a standard deviation of 1°. The probability (or confidence) that the friction angle is between 30° and 33° can be estimated as about 82%. This value is represented by the shaded area under the probability distribution in Figure A-1-17.

There are some philosophical differences between a confidence interval (a frequentist concept) and a credible interval (a Bayesian concept). A frequentist confidence interval would mean that 82% of the confidence intervals computed from repeated samples would contain the true value of the Phi angle. The true value is assumed to be a fixed, but unknown, value. A Bayesian credible interval would mean that there is an 82% probability that the parameter of interest (Phi angle) has a value between 30° and 33°. The true value is assumed to be a random variable. These differences are beyond the scope of this introductory manual and are typically inconsequential in practice.

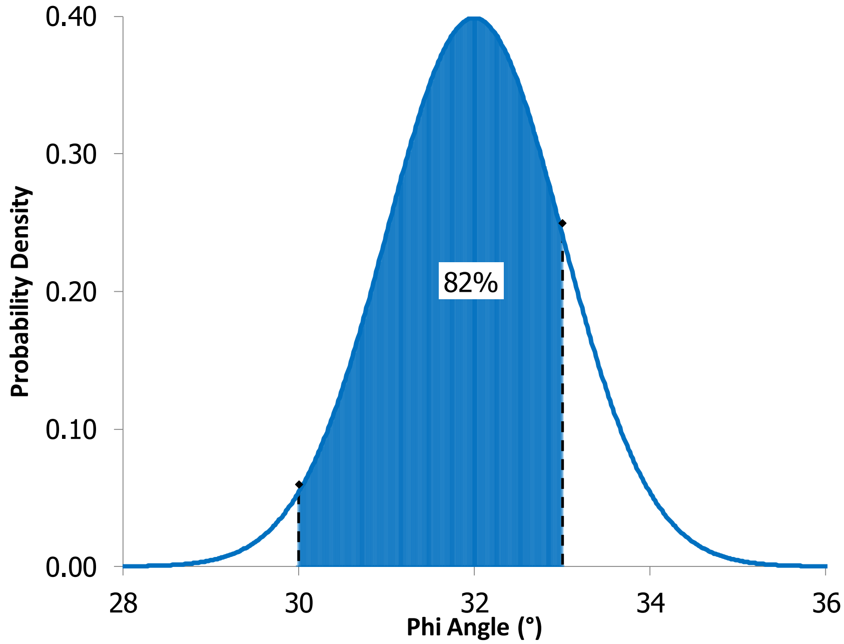

Figure A-1-17 Confidence Interval

A summary of confidence intervals associated with a normally distributed random variable is provided in

Table **A-1-4**. These confidence intervals are defined as a function of the mean and standard deviation.

Table A-1-4 Confidence Intervals for a Normal Distribution

Confidence	Interval	Meaning
68%	$\bar{x} - \sigma \leq x < \bar{x} + \sigma$	Plus or minus 1 standard deviation from the mean
95%	$\bar{x} - 2\sigma \leq x < \bar{x} + 2\sigma$	Plus or minus 2 standard deviations from the mean
99.7%	$\bar{x} - 3\sigma \leq x < \bar{x} + 3\sigma$	Plus or minus 3 standard deviations from the mean

A-1.7.5 Correlation

Correlation is the degree to which two or more variables are related to each other. For example, standard penetration test (SPT) blow counts might be correlated with the shear strength of soils. Higher blow counts might be an indicator of higher shear strengths. Correlation can be used to indirectly estimate parameters in a risk analysis. The concept can also be used to provide internal consistency between parameters within a risk model. For example, the effectiveness of a flood warning might be correlated with the time of day. Correlation alone does not imply or provide evidence of causation. A causal connection may exist only when there is a plausible cause and effect explanation.

A commonly used metric for the linear correlation between two variables is the Pearson product-moment correlation coefficient. Given a series of n measurements for X and Y, the sample correlation coefficient can be calculated using the following equation.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$

Equation A-1-24

This correlation coefficient can have a value between -1 and +1. Values closer to -1 and +1 suggest a stronger linear relationship. Values closer to zero suggest a weaker linear relationship. When using these types of metrics, a dose of caution is needed. Figure A-1-18 shows scatter plots for Anscombe's quartet. These are a set of four different sets of data pairs, with each set having the same values of mean, variance, correlation coefficient, and regression line, but very different appearance. They demonstrate that simple metrics may not always provide a sufficient basis for interpreting the data.

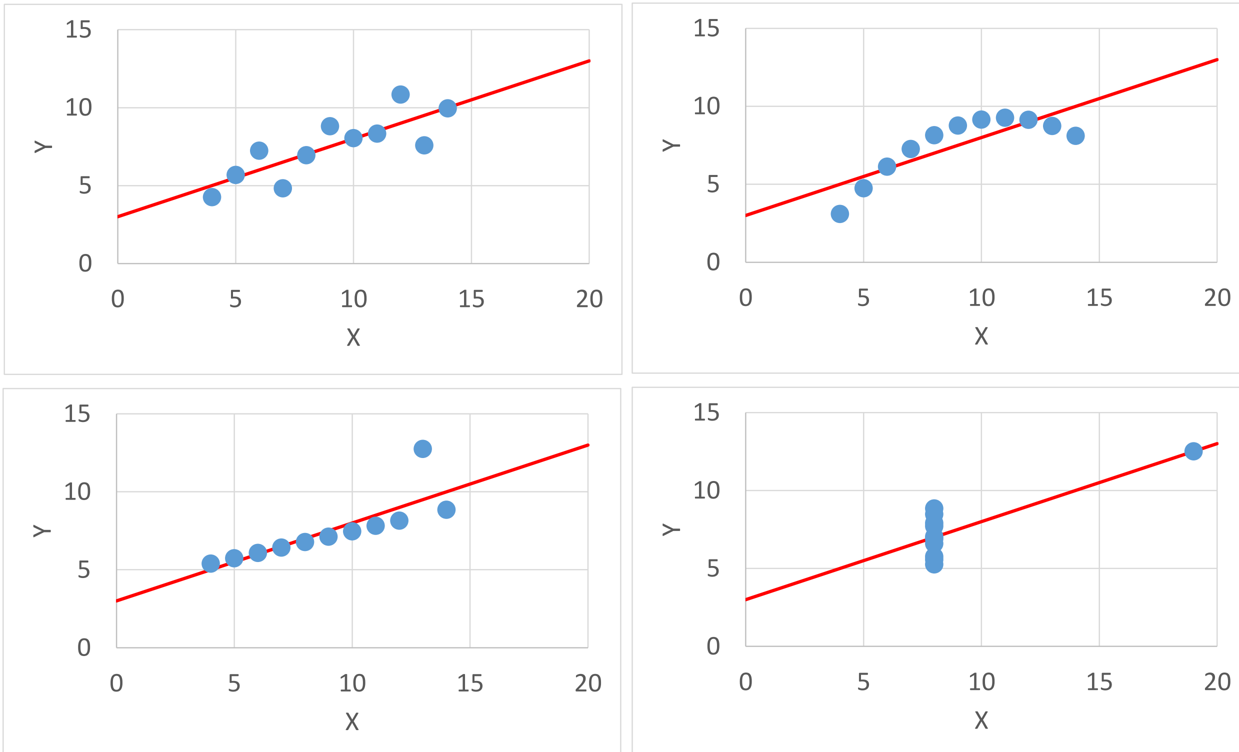

Figure A-1-18 Correlation Example

A-1.8 Bayesian Inference

Bayesian inference relies on Bayes' theorem to express the way in which a degree of belief probability should rationally change to account for new evidence. According to Ang and Tang (1975), the Bayesian method provides a useful approach when dealing with limited available information and when reliance on subjective judgments is necessary. It can be used to inform subjective judgments so that the available evidence is not given too much weight or too little weight when estimating probabilities.

The method begins with an estimate of the prior probability of an event based on available information. The significance of new information or evidence can then be considered by using Bayes' theorem to obtain an updated or posterior estimate of the event probability (Hartford and Baecher 2004). To illustrate the concept, it is convenient to start with the general form of Bayes theorem using the equation below, where $P(x|O)$ is the posterior probability of an event x given an observation O, $P(x)$ is the prior probability of the event x without the observation, $P(O|x)$ is the conditional probability of the observation O given the event x, and P(O) is the probability of the observation.

$$P(x|O) = \frac{P(x)P(O|x)}{P(O)}$$ **Equation A-1-25**

For example, let's assume a risk analyst estimates the initial probability for the presence of a permeable layer in a levee foundation to be P(x)=0.2, based on general knowledge of the regional geology; this is the prior probability. It was further judged that if a permeable layer did exist, its extent of the layer would be on the order of 200 feet. To improve the probability estimate, an exploration program is undertaken, with borings at 500-foot spacing along the levee alignment. No permeable layers are detected by the exploration program; this is the observation, O. Assuming that a 200-foot-long permeable layer does exist in a 500-foot reach, the length without the permeable layer would be 300 feet, so the probability of not finding it with a particular boring is P(O|x) = 300/500 = 0.6. This is the probability of the observation, O (no permeable layer detected), given that a permeable layer *does* exist (x). The total probability of not observing a permeable layer includes two possible events: the layer exists and it was not observed, or the layer does not exist. The probability that the layer *does not* exist is the complement of the prior probability estimate, 1-P(x)=0.8, and with no layer present, the probability of O (no layer detected) is 1.0. The total probability of not observing a permeable layer is therefore

$$P(O) = 0.2*0.6 + 0.8*1 = 0.92$$

Using Bayes' theorem, the updated or posterior probability of a permeable layer in the foundation following the exploration program can be calculated as shown below.

$$P(x|O) = \frac{0.2 * 0.6}{0.92} = 0.13$$

In this case, the observation O was consistent with the low prior probability, so it resulted in a fairly modest decrease.

A-1.9 Monte Carlo Analysis

Monte Carlo methods cover a broad range of computational algorithms that rely on repeated sampling to obtain numerical results. In risk analysis, Monte Carlo analysis is typically used to evaluate uncertainties when analytical solutions are difficult or do not exist. Common applications include stochastic modeling for hazards, reliability or limit state analysis for PFMs, stochastic modeling for consequences, and combining uncertainties for event trees analysis. For example, a simple model for estimating the average annual life loss (AALL) might be represented by the equation below.

$$AALL = P(Hazard) * P(Failure|Hazard) * Consequences|Hazard\ and\ Failure$$

When the model inputs are deterministic, a single value estimate for AALL is obtained as the product of the model inputs. When the model inputs are uncertain or random, the estimate of AALL is also uncertain or random. Monte Carlo analysis can be used to estimate the probability distribution of AALL.

The basic steps for any Monte Carlo analysis are listed below. These steps are applicable for use with any type of model. Specific applications for Monte Carlo analysis in dam and levee risk analysis are presented throughout the manual.

- Build a model (e.g. event tree calculation for AALL, limit state analysis for factor of safety, flood model for peak stage, etc)
- Assign probability distributions to the model inputs (e.g. uncertainty in branch probabilities, uncertainty in material properties, uncertainty in flood frequency)
- Define correlations among model inputs
- Sample the model inputs based on their distributions and correlations
- Run the model
- Record the model output
- Repeat for many samples of the model inputs
- Evaluate the probability distribution for the model outputs

A-1.10 References

Scriven, M. and Paul, R.W., 1987. Critical Thinking as Defined by the National Council for Excellence in Critical Thinking

Hartford, Desmond N.D. and Baecher, Gregory B, 2004. Risk and Uncertainty in Dam Safety

Vick, Steven G, 2002. Degrees of Belief – Subjective Probability and Engineering Judgment

Benjamin, Jack R. and Cornell, C. Allin, 1970. Probability, Statistics, and Decision for Civil Engineering

Ang, Alfredo H-S. and Tang, Wilson H., 1975. Probability Concepts in Engineering Planning and Design – Volume 1-Basic Principles

Ang, Alfredo H-S. and Tang, Wilson H., 2000. Probability Concepts in Engineering Planning and Design – Volume 2-Decision, Risk, and Reliability